Indexing and Classification at Northern Light

Presentation to CENDI Conference

“Controlled Vocabulary and the Internet”

Sept 29, 1999

Joyce Ward

Northern Light Technology, Inc.

www.northernlig
ht.com

NL's fundamental goals

- Combine Web data with quality information not on the Web ('Special Collection') in a single integrated search
- Make results set manageable for user (already a problem; worse after non-Web data is added)
- Take user from search ➔ full text in single session

www.northernlig
ht.com

Classification's fundamental goals

- Classify web to the same standard found for journal literature
- Develop subject, type, source, and language taxonomies to organize content regardless of source (NL Directory)
- Normalize all licensed taxonomies to NL Directory
- Present taxonomies in a way users can understand quickly

Gathering Web content

- The crawler (the robot Gulliver) discovers Web pages by following links & feeds them continuously to database
- Gulliver balances its time between crawling never-before-discovered pages, and updating pages it's already found
- Gulliver crawls randomly & in targeted fashion (as determined by librarian editors)
- Web database today includes about 178 million pages

www.northernlight.com

Indexing vs. classifying Web content

- Crawler sends pages to loader, which builds an index of every word on every page
- Loader sends pages to classifier, which attempts to determine what the page is *about,* what it *is,* where it is from, and the language it is written in
- Loader & classifier handle about 4 million pages/week

www.northernlig
ht.com

Gathering licensed content ('Special Collection')

- License full text from aggregators and publishers
- Use providers' metadata, when present, as basis for classification
- Special Collection includes about 20 million documents (compiling since 1995)

www.northernlig
ht.com

How classification is used

- All content is classified to subject, type, source, language taxonomies
- Engine uses this data to analyze & sort query results into Custom Search Folders™
- Displays prominent themes… “back of the book” index to your search results
- work with the user to refine the question (reference interview approach)

www.northernlig
ht.com

Power Search found **1,945 items** for:

"Senator Michael DeWine" or "Senato

SEARCH HELP

○ Special Collection ○ World Wide Web ◉ All Sources

Documents that best match your search

1. Senate starts to re-up Law Enforcement Block Grant. SPECIAL COLLECTION
   99% - Articles & General info: Senator Mike DeWine (R-Ohio) last week introduced legislation that would authorize the Local Law Enforcement Block Grant (LLEBG) program for the next five years and... 04/13/98
   Nation's Cities Weekly (news): Available at Northern Light

2. The Greater Columbus SuperSite : Government : Federal WWW
   98% - Directories & Lists: The Greater Columbus SuperSite : Government : Federal Government Government Columbus Franklin County Regional Counties Delaware Fairfield Licking Madison Pickaway Union Ohio Federal Military SuperSite... 09/22/98
   Non-profit sites: http:/ / www.columbus.org/ government/ federal.html

3. Battle of Fallen Timbers/Senator DeWine-Photos WWW
   98% - Articles & General info: Battle of Fallen Timbers/Senator DeWine-Photos Pictorial Sample of Activities and Artifacts DeWine Visit Photos: Maumee, Ohio The following pictures depict the visit of Ohio Sen.... Date Not Available
   Educational sites: http:/ / www.heidelberg.edu/ FallenTimbers/ FTDeWinePict.html

Narrow your search with Custom Search Folders™

Your search returned 1,945 items which we have organized into the following Custom Search Folders:

- Search Current News
- Special Collection documents
- Acquisitions & mergers
- Clinton, William J.
- Non-profit sites
- Elections
- Commercial sites
- Police science & law enforcement
- Immigration & emigration
- Drunk driving
- Holocaust
- Dayton Daily News
- Cincinnati Enquirer (news)
- all others...

www.northernlig
ht.com

How are folders used?

- To focus results on a specific aspect of of a topic
- To disambiguate queries

1,034,327 items found for:

what is balance? SEARCH HELP/HINTS

Documents that best match your search

Narrow your search with Custom Search Folders™

- Special Collection documents
- Commercial sites
- Sociology of the family
- Employee assistance programs
- Neurology
- Online banking
- Helicopters
- Martial arts
- Chinese philosophy
- all others...

1. WHAT IS BALANCE?

84% - Articles & General info: WHAT IS BALANCE? Back to New Evangelicanism Reports. Back to the Way of Life Home Page Way of Life Literature Online Catalog You Can Own...11/09/97

Personal Page: http://www.dsinclair.com /~dcloud/fbns /whatisbalance.htm

2. Emotional Stability is Balance

77% - Articles & General info: Emotional Stability is Balance Emotional Stability is Balance - 1 He is unbalanced - 2 She's not on an even keel - 3 They're upset...

03/24/95

Educational site:http://cogsci.berkeley.edu/metaphors/ EmotionalStabilityIsBalance.html

3. What is balance?

73% - Biographical sources: "What is balance?" This is an ongoing, soul- searching, head-scratching question that my husband, Don, and I ponder on a regular bases....07/01/96

Exceptional parent (magazine): Available at Northern Light

www.northernlight.com

How are folders used?

- To focus results on a specific aspect of of a topic
- To disambiguate queries
- To answer questions directly

Buy books at
BARNES&NOBLE

Power Search found **255 items** for:

Who said "Death, be not proud"?

SEARCH HELP

○ Special Collection ○ World Wide Web ◉ All Sources

Documents that best match your search

1. Death, be not proud
   82% - Articles & General info: In the past two weeks, three Tulsa teen-agers from different schools have died, sending ripples of shock through Tulsa. Kevin Godwin, a freshman at Central... 02/12/99
   Tulsa World - Tulsa, OK (news): Available at Northern Light
   SPECIAL COLLECTION

2. Death Be Not Proud
   81% - Articles & General info: Death Be Not Proud Based on a poem by John Donne Dark room, dim lights, and an old, dust covered chair, planted in the middle,... 02/21/99
   Commercial site: http:/ / www.asylumbbs.com/ mojo/ death.txt
   WWW

3. Death, Be Not Proud
   80% - Articles & General info: HER-2: The Making of Herceptin, A Revolutionary Treatment for Breast Cancer, by Robert Bazell. Random House, 214 pp., $23.95. IT IS A SCENE every woman... 10/11/98
   Newsday: Available at Northern Light
   SPECIAL COLLECTION

Narrow your search with Custom Search Folders™

Your search returned 255 items which we have organized into the following Custom Search Folders:

- Search Current News
- Poetry
- Donne, John - works, studies & biography
- Commercial sites
- Theater & Performing arts
- Cancer diagnosis & treatment
- Euthanasia
- Cancer

www.northernlight.com

Subject classifying the Web

- Manual approaches do not scale: cost of classifying 1 journal article=$1.70. Multiplied by 178 million web pages = about $300 million
- Automatically determine document's subject, type, source and language metadata
- Artificial intelligence system uses controlled vocabulary to classify pages

www.northernlight.com

Automatic classification techniques

- Mixed (vs totally manual, totally automatic): human-directed
- Based on words contained in document
- Uses Term Frequency/Inverse Document Frequency methods to match document to term(s) from controlled vocabulary
- Each term has set of co-occurring terms derived from training set
- Document must have a strong degree of 'aboutness' to class

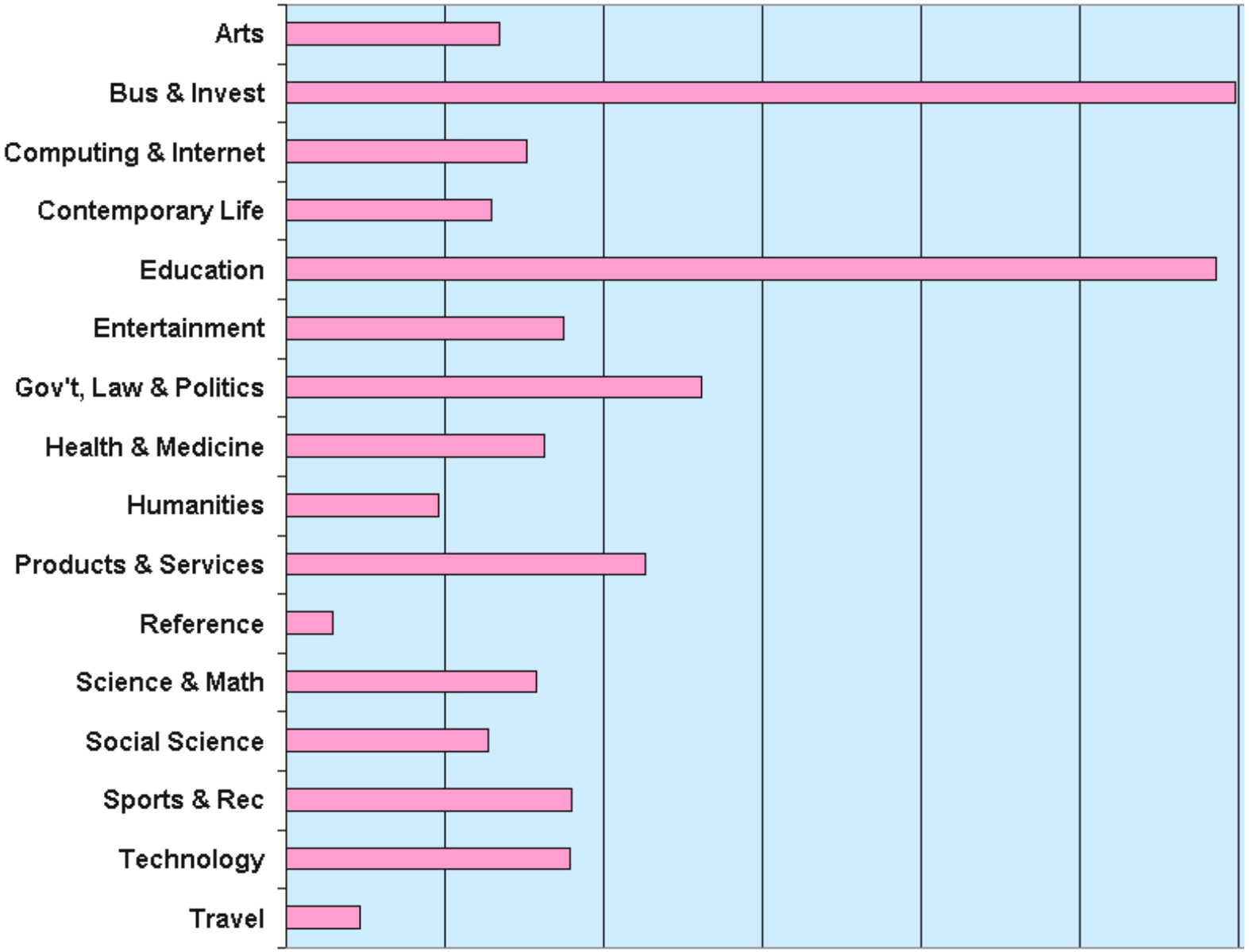
www.northernlight.com

NL's subject vocabulary

- Subject scope is unlimited (as in LC, Dewey, Yahoo)
- Major points of reference were DDC, LC Subject headings, UMI subject headings, and subject-specialized classification schemes
- Unique, selective conflation of these
- Mapping NL with content partners' vocabularies gives freshness, completion
- 25,000 concepts; 200-300,000 concept equivalents
- 16 top-level subjects; hierarchies 7 - 9 levels

www.northernlig

NL Subject areas and relative size



Why bother classifying? why not use contents of <meta> tags?

- Metadata is present in
  - less than 30% of web pages (Site Metrics, 97 & 98)
  - slightly more than 40% of web pages (NL sample, Oct 98)
- Most of that is generated by page creation software & carries no ‘subject’ freight
- Subject metadata as provided by page creators is mostly spam
- Trace amounts of well-formed metadata on the web at this time

www.northernlig
ht.com

Subject <meta> from a randomly crawled page

- naples.net:

"games,games,games,gamez,gamez,game,game,game,gamez,nes,nes,nes,snes,snes,snes,sega,sega,sega,genesis,genesis,genesis,roms,roms,roms,emulator,emulator,emulator,emulators,emulators,emulators,shareware,shareware,shareware,download,download,download,games,games,games,gamez,gamez,game,game,game,gamez,nes,nes,nes,snes,snes,snes,sega,sega,sega,genesis,genesis,genesis,roms,roms,roms,emulator,emulator,emulator,emulators,emulators,emulators,download,download,download,games,games,games,gamez,gamez,game,game,game,gamez,nes,nes,nes,snes,snes,snes,sega,sega,sega,genesis,genesis,genesis,roms,roms,roms,emulator,emulator,emulator,emulators,emulators,emulators,download,download,download,games,games,games,gamez,gamez,game,game,game,gamez,nes,nes,nes,snes,snes,snes,sega,sega,sega,genesis,genesis,genesis,roms,roms,roms,emulator,emulator,emulator,emulators,emulators,emulators,download,download,download"

www.northernlight.com

Subject classifying the Special Collection

- Map the information provider's metadata to the NL Directory
- Extend NL Directory where necessary
- Automatically classify where metadata is non-existent or when fewer than 2 subjects are provided
- All synonyms are preserved & used to automatically match new vocabs to NL Directory

www.northernlight.com

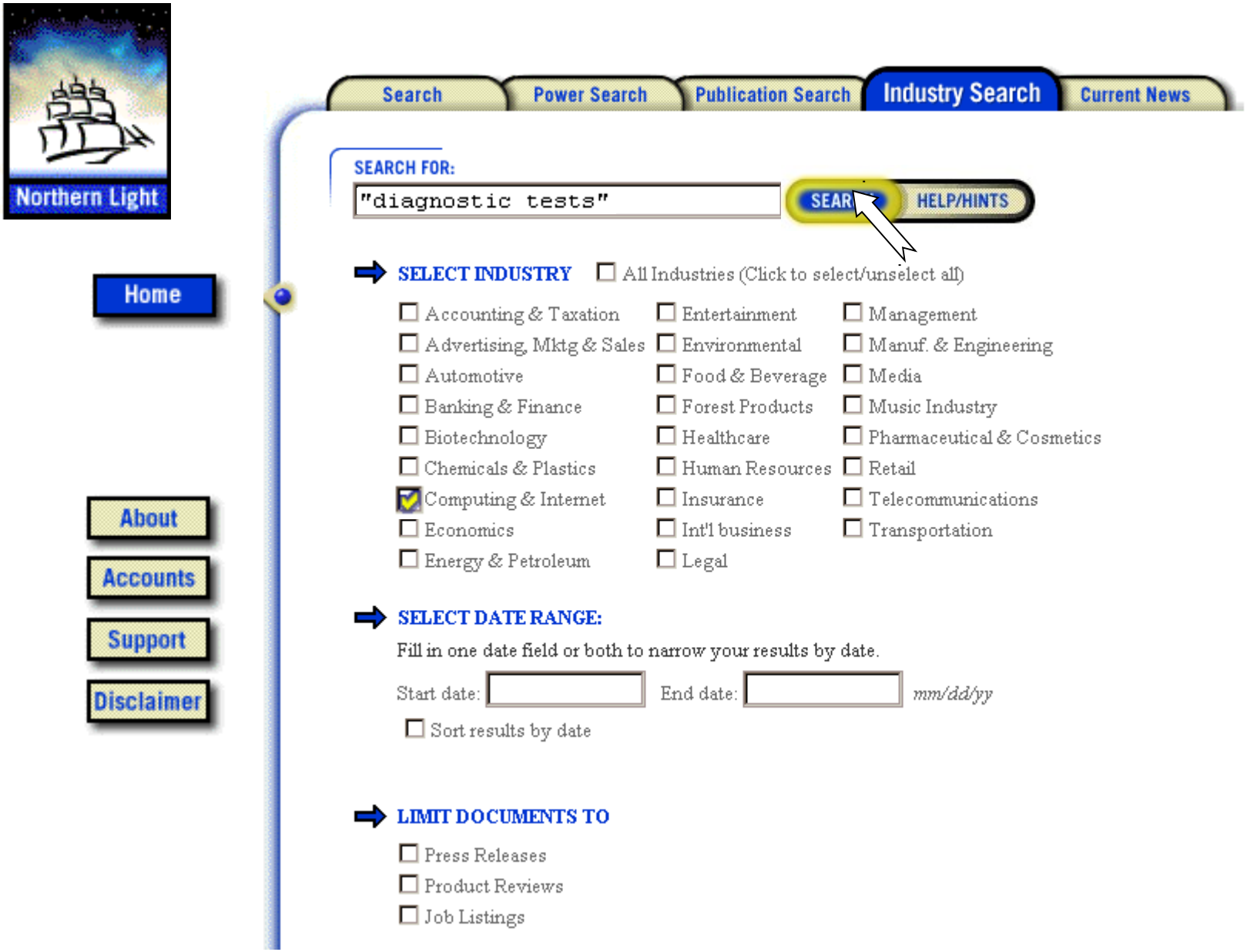
Mapping FDCH categories to NL

FDCH Category	NL Subject	Subject/Type/Region NEE
Birth control	172	Contraception
Bombings	15778	Terrorism
Budget	39605	Government finance
Business	88	Business & Investing
Cancer	10660	Cancer
Capital punishment	15679	Death penalty
Charity	6136	Charities & Foundations
Chemicals	4643	Chemical products
Children	6756	Childhood
Cities	16850	Urban planning
Civil rights	150	Civil rights & discrimination

Controlled vocabularies enable specialized search engines

- Vocabularies can be used as powerful subject filters

www.northernlig
ht.com



www.northernlig
ht.com

Industry Search found **646 items** for:

"diagnostic tests" SEARCH HELP/HINTS

Documents that best match your search

1. InfoBase: Ledcodes (12/29/97)
   99% - Directories & Lists: InfoBase: Ledcodes (12/29/97) Network Hardware Support I n f o r m a t i o n B a s e Product: 6611 Document: ER0144X... 03/31/98 WWW
   Commercial site: http://www.networking.ibm.com /nes/nesibase/nhdinfo /251a_4a2.htm

2. Diagnostic Tests
   86% - Articles & General info: Diagnostic Tests CalComp Tablets ccmouse.... 09/09/98 WWW
   Commercial site: http://www.calcomp.com /a_station /as_diags.htm

3. FAQ: How do I run diagnostic tests using DataSMART Installer software?
   79% - Articles & General info: FAQ: How do I run diagnostic tests using DataSMART Installer software? DSU/CSU Magnum PACESETTER How do I run diagnostic tests using the DataSMART Installer? You... 08/07/98 WWW
   Commercial site: http://www.kentrox.com /support /faqs/dsucsu/installer /5.html

4. Cisco - FRP/FRM Background Diagnostic Test Issues
   79% - Articles & General info: Cisco - FRP/FRM Background Diagnostic Test Issues Field Notice: FRP/FRM Background Diagnostic Test Issues Device Code Channel Address Conflicts February 7, 1997 Problem Description... 06/04/98 WWW
   Commercial site: http://cio.cisco.com /warp/public /770/25.html

Narrow your search with

Custom Search Folders™

- Search Current News
- Special Collection
- Computer networks
- Local area networks
- Modems
- Cable modems
- Personal computers
- Computer caches
- Buses (computer)
- Health care software
- Software industry
- Circuit design
- all others...

Industry Search found 646 items for:

"diagnostic tests" SEARCH HELP/HINTS

Documents that best match your search

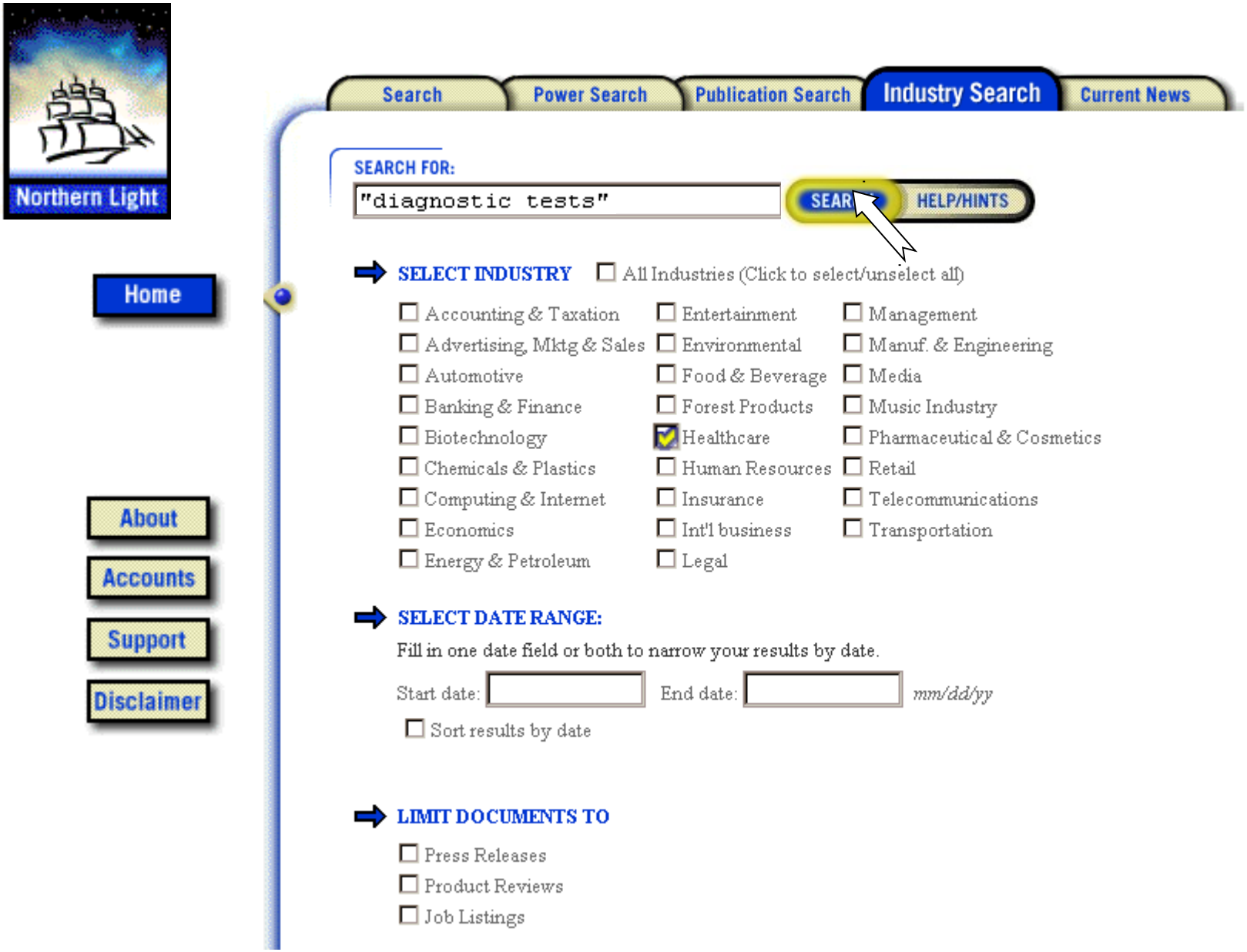
1. InfoBase: Ledcodes (12/29/97)
   99% - Directories & Lists: InfoBase: Ledcodes (12/29/97) Network Hardware Support I n f o r m a t i o n B a s e Product: 6611 Document: ER0144X... 03/31/98 WWW
   Commercial site: http://www.networking.ibm.com /nes/nesibase/nhdinfo /251a_4a2.htm

2. Diagnostic Tests
   96% - Articles & General info: Diagnostic Tests CalComp Tablets ccmouse.... 09/09/98 WWW
   Commercial site: http://www.calcomp.com /a_station /as_diags.htm

3. FAQ: How do I run diagnostic tests using DataSMART Installer software?
   79% - Articles & General info: FAQ: How do I run diagnostic tests using DataSMART Installer software? DSU/CSU Magnum PACESETTER How do I run diagnostic tests using the DataSMART Installer? You... 08/07/98 WWW
   Commercial site: http://www.kentrox.com /support /faqs/dsucsu/installer /5.html

4. Cisco - FRP/FRM Background Diagnostic Test Issues
   79% - Articles & General info: Cisco - FRP/FRM Background Diagnostic Test Issues Field Notice: FRP/FRM Background Diagnostic Test Issues Device Code Channel Address Conflicts February 7, 1997 Problem Description... 06/04/98 WWW
   Commercial site: http://cio.cisco.com /warp/public /770/25.html

www.northernlig
ht.com



www.northernlig
ht.com

Industry Search found **4,650 items** for:

"diagnostic tests" SEARCH HELP/HINTS

Documents that best match your search

1. [Selling] - Medical Diagnostic Test Kits
   94% - Directories & Lists: [Selling] - Medical Diagnostic Test Kits 2/18/98 Medical Diagnostic Test Kits Message We are offering various types of diagnostic test kits for home... Date Not Available
   Commercial site: http://www.asiapacific.com /infocenter /selling /s-feb18-11.html
   WWW

2. Chapter 5.1, Table 2 - Fever evaluation - laboratory and diagnostic tests
   92% - Articles & General info: Chapter 5.1, Table 2 - Fever evaluation - laboratory and diagnostic tests Chapter 5.1, Table 2 Fever evaluation - laboratory and diagnostic tests CD4 count... 03/20/97
   Educational site: http://hivinsite.ucsf.edu /akb /1994/5-1/toctable2.html
   WWW

3. Heart Diagnostic Tests and Procedures
   88% - Articles & General info: Heart Diagnostic Tests and Procedures Heart Diagnostic Tests and Procedures Heart Damage Detection Tests to Diagnose Heart Disease Outpatient Cardiac Catheterization Ambulatory Electrocardiogra... 08/27/98
   Non-profit site: http://www.americanheart.org /Heart /Heart_Attack /Diagnostic_Tests /index.html
   WWW

4. Interpretation of Diagnostic Tests
   88% - Articles & General info: Interpretation of Diagnostic Tests Lippincott-Raven... Interpretation of Diagnostic Tests Sixth Edition By Jacques Wallach, MD This comprehensive and user-friendly electronic reference is ideal for all... 06/18/98
   WWW

www.northernlight.com

Narrow your search with
Custom Search Folders™

- Search Current News
- Special Collection
- Pharmaceuticals industry
- Diagnostic test agents
- Pharmacists & pharmacy services
- HIV test
- Genetics
- Patent law
- Heart (Physiology)
- Allergies
- Orthopedic surgeons
- Alzheimer's disease
- Penicillin
- all others...

Industry Search found 4,650 items for:

"diagnostic tests" SEARCH HELP/HINTS

Documents that best match your search

1. [Selling] - Medical Diagnostic Test Kits
   94% - Directories & Lists: [Selling] - Medical Diagnostic Test Kits 2/18/98 Medical Diagnostic Test Kits Message We are offering various types of diagnostic test kits for home... Date Not Available
   Commercial site: http://www.asiapacific.com /infocenter /selling /s-feb18-11.html

2. Chapter 5.1, Table 2 - Fever evaluation - laboratory and diagnostic tests
   92% - Articles & General info: Chapter 5.1, Table 2 - Fever evaluation - laboratory and diagnostic tests Chapter 5.1, Table 2 Fever evaluation - laboratory and diagnostic tests CD4 count... 03/20/97
   Educational site: http://hivinsite.ucsf.edu /akb /1994/5-1/toctable2.html

3. Heart Diagnostic Tests and Procedures
   88% - Articles & General info: Heart Diagnostic Tests and Procedures Heart Diagnostic Tests and Procedures Heart Damage Detection Tests to Diagnose Heart Disease Outpatient Cardiac Catheterization Ambulatory Electrocardiogra... 03/27/98
   Non-profit site: http://www.americanheart.org /Heart /Heart_Attack /Diagnostic_Tests /index.html

4. Interpretation of Diagnostic Tests
   88% - Articles & General info: Interpretation of Diagnostic Tests Lippincott-Raven... Interpretation of Diagnostic Tests Sixth Edition By Jacques Wallach, MD This comprehensive and user-friendly electronic reference is ideal for all... 06/18/98

www.northernlight.com

Are controlled vocabularies important in the Web environment?

- At Northern Light, they are essential to the way we organize results for users
- They provide a unified view of all content, regardless of source
- They enable creation of specialized (‘vertical’) search products

www.northernlig
ht.com

Joyce Ward

VP, Editorial Services

jward@northernlight.com

www.northernlig
ht.com